Overview

Workflow

Accuracy

TSEBRA: Transcript Selector for BRAKER

BRAKER-TSEBRA-Workshop
December 15th 2022

Lars Gabriel
Katharina J. Hoff
Tomáš Brůna
Mark Borodovsky
Mario Stanke

Presenting author e-mail: lars.gabriel@uni-greifswald.de

TSEBRA - Overview

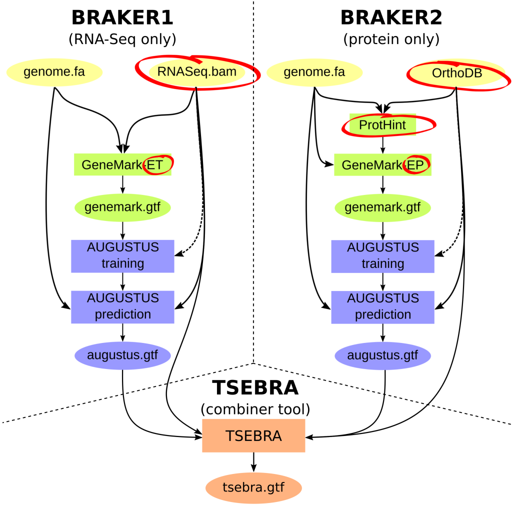
Task

Create a BRAKER annotation based on RNA-Seq and protein evidence by combining a BRAKER1 and a BRAKER2 prediction.

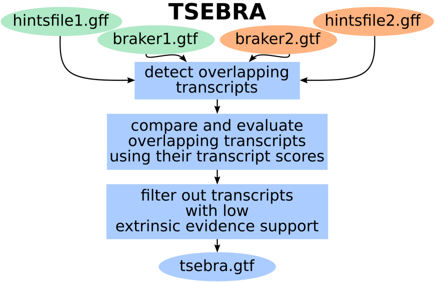
Input

- One or more gene sets (e.g. `augustus.hints.gtf` from BRAKER)
- Hints from extrinsic evidence as intron or start/stop codon positions (e.g. `hintsfile.gff` from BRAKER)
- Configuration file with parameters (e.g. `default.cfg`)

Hoff et al. 2016. *Bioinformatics*. 32(5):767–9.
Brůna, Hoff et al. 2021. *NAR Genomics and Bioinform.* 3(1):lqaa108.
Gabriel et al. 2021. *BMC Bioinformatics*. 22: 566.

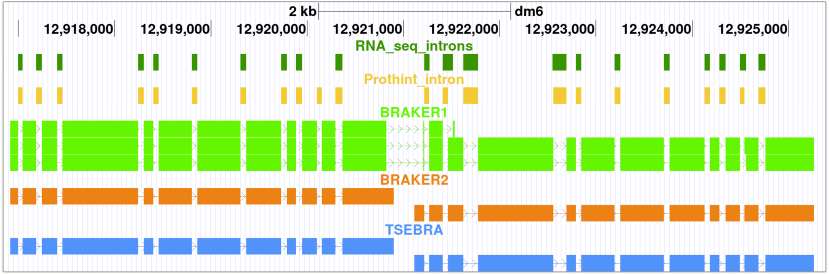


TSEBRA - Workflow


TSEBRA
hintsfile1.gff
braker1.gtf
braker2.gtf
hintsfile2.gff
detect overlapping transcripts
compare and evaluate overlapping transcripts using their transcript scores
filter out transcripts with low extrinsic evidence support
tsebra.gtf

Transcript scores

1. Percentage of introns supported by at least one hint.
2. Number of hints supporting any intron.
3. Percentage of start/stop codons supported by hints.
4. Number of hints supporting start/stop codons.



TSEBRA - Configuration File

```
# Weight for each hint source
# Values have to be >= 0
P 0.1
E 10
C 5
M 1
# Required fraction of supported introns
# or supported start/stop-codons for a transcript

intron_support 0.75
stasto_support 1
# Allowed difference for each feature
# Values have to be in [0,1]
e_1 0
e_2 0.5
# Values have to be >0
e_3 25
e_4 10
```

P: protein hints

E: RNA-Seq hints

C: hints from proteins with high alignment scores

M: manual hints

sets the weight of each hint source for the transcript scores

filter out transcript if
- <75% of introns and
- not start and stop codon supported

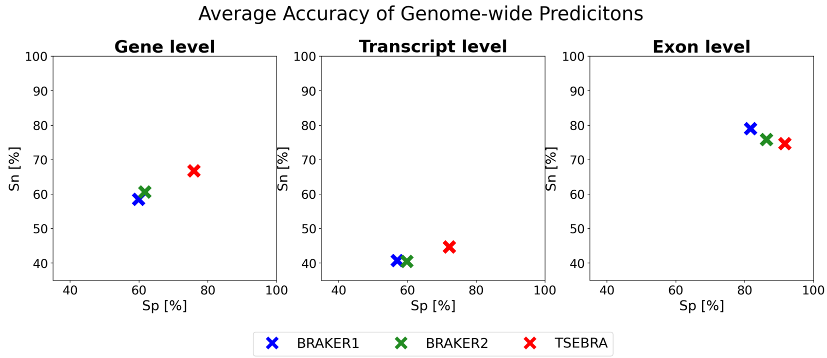
determines how strictly TSEBRA filters out transcripts with low evidence support

thresholds used for the pairwise comparison of transcript scores, one for each score

Transcript scores

1. Percentage of introns supported by at least one hint.
2. Number of hints supporting any intron.
3. Percentage of start/stop codons supported by hints.
4. Number of hints supporting start/stop codons.

TSEBRA - Accuracy



Species: *Arabidopsis thaliana*, *Caenorhabditis elegans*, and *Drosophila melanogaster*.

Extrinsic evidence:

- paired RNA-Seq short reads
- large protein database including distantly related species

Acknowledgements

Funding

This research is supported by US National Institutes of Health grant GM128145 to Mark Borodovsky and Mario Stanke.

Co-Authors

Katharina J. Hoff
Tomáš Brůna
Alexandre Lomsadze
Mark Borodovsky
Mario Stanke

Availability

- `https://github.com/Gaius-Augustus/TSEBRA/`

References

Lomsadze et al. ”Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm.” Nucleic acids research 42.15 (2014): e119-e119.

Brůna et al. ”GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins.” NAR genomics and bioinformatics 2.2 (2020): lqaa026.

Stanke et al. ”Using native and syntenically mapped cDNA alignments to improve de novo gene finding.” Bioinformatics 24.5 (2008): 637-644.

Hoff et al. ”BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS.” Bioinformatics 32.5 (2016): 767-769.

Brůna et al. ”BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database.” NAR genomics and bioinformatics 3.1 (2021): lqaa108.

Gabriel et al. ”TSEBRA: Transcript Selector for BRAKER.” BMC Bioinformatics 22: 566 (2021).

Evgenia et al. ”OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs.” Nucleic Acids Research 47.D1 (2019): D807–D811.